The Transmission of Scientific Knowledge to Wikipedia

Misha Teplitskiy
Dept. of Sociology and KnowledgeLab
University of Chicago
misha@uchicago.edu

Grace Lu
KnowledgeLab
University of Chicago
gracelu@uchicago.edu

Eamon Duede
KnowledgeLab
University of Chicago
eduede@uchicago.edu

Credits

This research was made possible by a grant (#39147) to the Metaknowledge Network by the John Templeton Foundation.

Abstract

This paper compares the scientific literature used most often by scientists to the scientific literature referenced on the English-language Wikipedia. Previous studies have raised concerns that editors of science-related articles on Wikipedia are biased toward easily available sources and underrepresent particular scientific fields. Most often, these studies examine references on Wikipedia only but make claims about how well or poorly Wikipedia represents the scientific literature as a whole. In contrast, the present study begins with the scientific literature. We use the *Scopus* database to identify the 250 most heavily used journals in each of 26 research fields (4620 journals in total), and estimate a variety of models to identify what makes these journals more or less likely to be cited on Wikipedia. We find that, controlling for impact factor and open access policy, Wikipedia over-represents journals from the *Social Sciences*, and under-represents journals from the *Physical Sciences* and *Health Sciences*. An open-access policy is not associated with increased Wikipedia presence. The most significant predictor that a journal will be cited on Wikipedia is its impact factor.

Introduction

Wikipedia has become a top destination for information of all kinds, including information about science (Spoerri, 2007). According to Alexa Internet, a web traffic analytics firm, Wikipedia is the 6th most visited website in the United States, and the 7th world wide ("Traffic"). Given that so many people rely on Wikipedia as a source of information about science, it is reasonable to ask whether and to what extent the science that is written about on Wikipedia is (1) an accurate representation of the knowledge within the academic literature and is (2) of sufficient quality.

Previous studies have raised concerns about both representation (Samoilenko and Yasseri, 2014) and quality (see Mesgari et al., 2015 for a recent review). While Wikipedia is an open source and collaborative effort, a vanishingly small number of its overall users actually contribute content and edits ("Trends"). Editors' demographics may bias them in favor of "topics that interest the young and Internet-savvy" (Denning et al., 2005). Additionally, one of Wikipedia's basic tenets is that its entries should be based on reliable and easily verifiable sources. To that end, Wikipedia's "Core Content Policy" on "Verifiability" provides guidelines which state that "where available, academic and peer-reviewed publications are usually the most reliable sources, such as in history, medicine, and science" "Verifiability"). While the peer review system as a whole is not without its share of problems, it is almost certainly the case that placing a premium on publications that have gone through the peer-review process does much to establish the reliability of Wikipedia's entries (Lucy Holman Rector, 2008). Yet, access to the vast majority of reliable, peer-reviewed scientific literature is restricted to holders of expensive subscriptions (Björk and Solomon, 2012), thereby creating a tension between Wikipedia's goals of making entries reliable on the one hand, and verifiable on the other. This paper seeks, in part, to understand how this tension is resolved in practice. For instance, some have argued that Wikipedia's editors cannot fully resolve this tension, and simply rely on references that are low quality, public, open access, or produced by partisan sources (Ford et al., 2013; Luyt and Tan, 2010). Some studies (Nielsen, 2007; Shuai et al., 2013) do find positive correlations between academic citations and Wikipedia mentions, but do not take into account accessibility.

This paper contributes to the existing literature by systematically comparing the literature relied on most by scientists to that cited on Wikipedia. We use the *Scopus* database to identify a large sample of the most important journals within each scientific field and use Wikipedia's structured reference tags to identify all citations to scientific journals. This design allows us to update earlier studies that have examined variation among the journals cited on

Wikipedia (Nielsen, 2007) and, crucially, examine which journals are *not* cited.

Our early results from this on-going research concern the role of a journal's impact factor, open access, and topic on representation in Wikipedia. We find that, controlling for impact factor and open access policy, Wikipedia over-represents journals from the *Social Sciences*, and under-represents journals from the *Physical Sciences* and *Health Sciences*. An open-access policy is not associated with increased Wikipedia representation. The most significant predictor that a journal will be cited on Wikipedia is its impact factor.

These findings should be interpreted with care because it is unclear whether Wikipedia or the academic literature should be taken as the golden standard of impact. This consideration will be elaborated in the conclusion.

Data and Methods

Data sample

Indexing over 21,000 peer-reviewed journals, with more than 2,800 classified as open access, *Scopus* is the world's largest database of scientific literature ("Scopus"). We obtained metadata on the 250 highest-impact journals within each of the following 26 sub-categories: *Agricultural Sciences, Arts and Humanities, Biochemistry and General Microbiology, Business Management and Accounting, Chemical Engineering, Chemistry, Computer Science, Decision Sciences, Earth and Planetary Sciences, Economics and Finance, Energy Sciences, Engineering, Environmental Sciences, Immunology and Microbiology, Materials Sciences, Mathematics, Medicine, Neurosciences, Nursing, Pharmacology, Physics, Psychology, Social Science, Veterinary Science, Dental, Health Professions*. These sub-categories were nested under the following "top level" categories: *health sciences, life sciences, physical sciences, and social sciences*.

Impact factor was measured by the 2013 SCImago Journal Rank (SJR) impact factor[1]; each journal's metadata included this impact factor, top-level- and sub- categories, number of articles published, and open-access policy.

Data on Wikipedia sources was obtained from a 2014-11-15 database dump of the full English-language Wikipedia. We parsed every page, and extracted all references to science that use Wikipedia's standardized *journal* attribute within references that use the *cite* tag. In all, there were 106,772 references to the scientific literature with 10,622 unique "journals" represented in Wikipedia. In many cases the "journal" cited was not an academic journal but a blog, non-academic website, or newspaper. The efforts to match the various "journal" strings to *Scopus* metadata, discussed below, were thus necessarily imperfect.

Disambiguation, data cleaning

We checked each of the referenced journal names on Wikipedia against a list of common *ISI* journal name abbreviations and, additionally, converted all abbreviated titles to canonical form.

Many of the 250 top journals in a given *Scopus* category were also in the top 250 of another category. The list of candidate journals was thus less than 250 * (number of research fields). We also removed from the data sample those journals that appeared in our data as having published no more than 100 articles. These cases were most often journals that have actually published many more articles but were indexed by Scopus relatively recently.

The final data consisted of 4620 unique journals, 307 of which are categorized by *Scopus*[2] as "open access", and 1779 of which do not appear in Wikipedia at all. Starting instead with Wikipedia, 55,267 of its 106,722 (51.7%) scientific references were linkable to Scopus. The precise composition of the remaining 51,505 references is unclear, but as stated previously, it includes a very large number of non-scientific resources (e.g. New York Times).

percent_cited and Other Variables

The quantity we sought to explain is *percent_cited* -- the *percent of a journal's articles that are cited on Wikipedia*. We chose this measure for two reasons. First, the raw number of articles journals publish varies tremendously. For example, the journal *PLoS One* has published more than 100,000 articles in a little over 8 years ("PLoSOne"), while the *American Journal of Sociology* has published about 10,000 articles in a little over 100 years. Focusing on raw citation counts in Wikipedia would privilege large-volume journals like *PLoS One*. On the other hand, *percent_cited* is normalized against a journal's output, so that journals may be compared on their topics, accessibility, and prestige, not simply on size. Second, the journal as the unit of analysis (instead of article-level analysis) greatly simplified disambiguation and matching Wikipedia references to *Scopus* metadata. It should be noted that while the article as the unit of analysis may appear preferable due to the ability to judge its impact in science via citations, a journal's meta-data captures this same citation-based impact, albeit more coarsely, via the impact factor. Figure 1 illustrates the distribution (kde) of *percent_cited* with logged x-axis.

[1] "SCImago Journal Rank (SJR) is a measure of scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from." (http://en.wikipedia.org/wiki/SCImago Journal Rank). It is especially useful for comparing journals across research areas.

[2] *Scopus* designations of open access are based on the Directory of Open Access Journals (www.doaj.org).

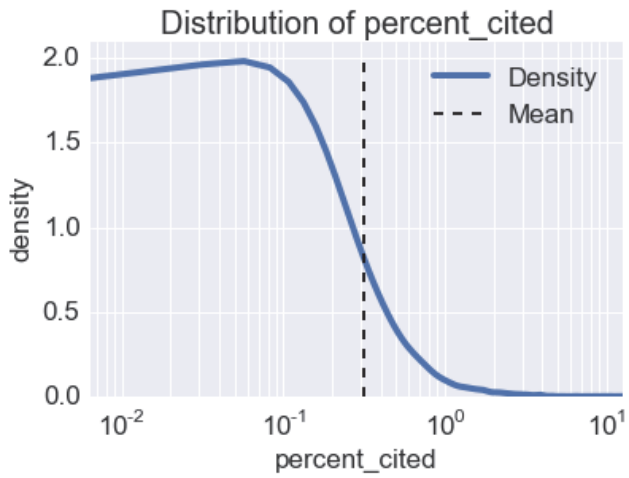

Figure 1. Distribution (kde) of percent_cited. *X-axis is logged.*

The distribution of impact factor (*SJR2013*) is illustrated in Figure 2.

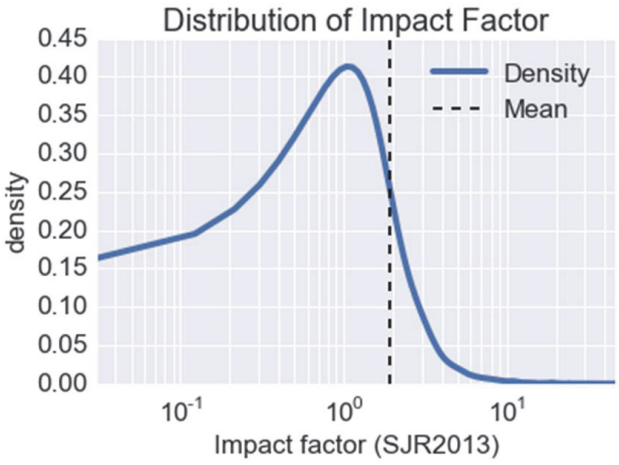

Figure 2. Distribution (kde) of impact_factor. *X-axis is logged.*

Both *percent_cited* and *impact factor* are highly skewed, so they were ln-transformed in the analyses that follow. We used a generalized linear model[3] with binomial probability distribution and a logit link function to model how *percent_cited* is associated with the following explanatory variables: the *journal's (SJR) impact factor*, its *open-access* policy, and both its *field* and *subfield*. Table 1 describes these variables.

[3] A logistic regression model produced qualitatively identical results. We followed (Baum, 2008) in using the generalized linear model with the stated parameters.

Variable name	*Valid Observations*	*Mean*	*Std.*	*Min*	*Max*
percent_cited	4634	0.32	0.64	0	12.7
impact factor	4585	1.90	2.48	0.10	45.9
open access	4634	6.7% O.A.	----	0	1
phys. sci.	4634	31.6%	---	0	1
life sci.	4634	42.9%	---	0	1
health sci.	4634	32.4%	---	0	1
soc. sci.	4634	29%	---	0	1

Table 1. Descriptive statistics of variables. Note that many journals fall under several top- and sub-level research areas.

The 26 subcategories in the Subcategory Analysis were similarly represented by dummy variables.

Results

First, we present scatter plots of *percent_cited* vs. *impact factor* and then analyze this relationship with a generalized linear model, separately for top-level and sub-categories. Figure 3 illustrates the scatter plot of log *percent_cited* and *impact factor*.

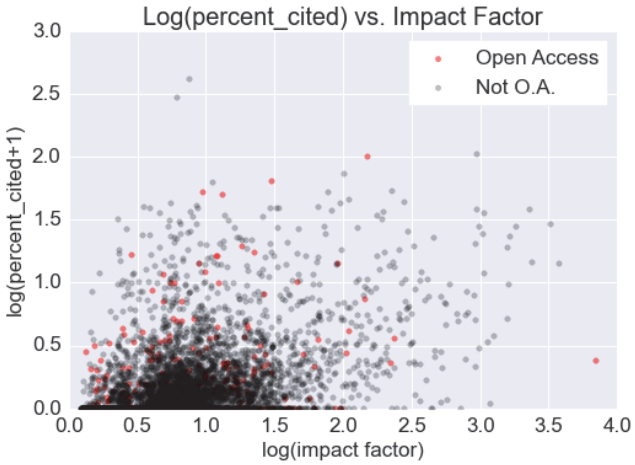

Figure 3. percent_cited vs. impact factor. Open access journals are colored red.

Figure 3 suggests that there is indeed the expected correlation between *percent_cited* and impact factor.

The relationships between *percent_cited* and the various categorical variables may be explored with boxplots. Figures 4 and 5 present boxplots of *percent_cited* grouped by top-level category and open access policy, respectively.

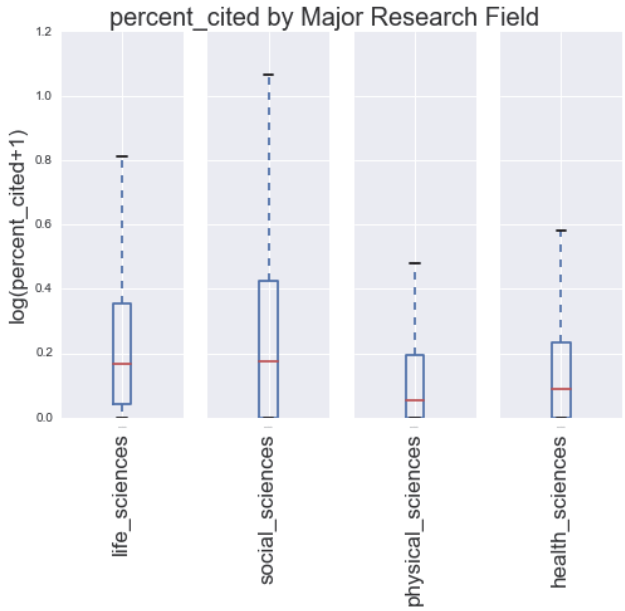

Figure 4. Boxplot of percent_cited by top-level research category.

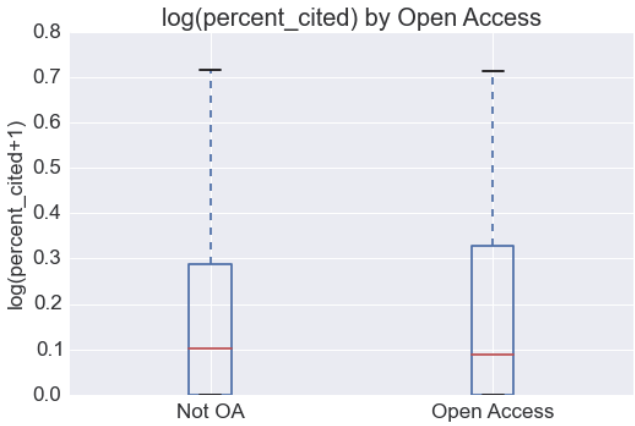

Figure 5. Boxplot of percent_cited by open-access policy.

Figure 4 suggests that there may be slight differences in *percent_cited* across topical categories, with life and social sciences being better represented than the physical sciences, and health sciences in the middle. On the other hand, open access does not appear to correlate with *percent_cited*. To explore these relationships statistically we report results from a generalized linear model.

Top-level categories

Here we present results of the model fit using *Scopus'* 4 top-level subject categories: *health sciences, life sciences, physical sciences, and social sciences*. Table 2 contains the coefficients indicating how these subject categories, along with a journal's open access policy (open or closed), and (logged) SJR impact factor are associated with the percent of its articles that are cited in Wikipedia, (logged) *percent_cited*.

Variable	*coefficient*	*std. err.*	*P > |t|*
Open Access=True	0.34	0.39	0.382
Log(Impact Factor)	**0.87**	**0.151**	**0.000**
Physical Sciences	**-0.48**	**0.261**	**0.063**
Social Sciences	**.54**	**0.269**	**0.046**
Life Sciences	0.28	0.260	0.285
Health Sciences	**-0.48**	**0.268**	**0.072**

Table 2. percent_cited *vs Predictors, (Predictors statistically significant at the 0.1 level are bolded. The intercept is not reported.)*

The table indicates that, controlling for impact factor and open-access policy, journals from the *Social Sciences* are likelier to be cited on Wikipedia than journals from the *Physical* and *Health Sciences*. Contrary to our expectations, an open access policy is not significantly associated with *percent_cited*.

The journal-level feature most significantly associated with the journal's representation in Wikipedia is its impact factor. The size of this effect is more easily interpreted using un-logged impact factor and *percent_cited*; using this alternative specification in an OLS model, a one unit increase in impact factor is associated with 0.083% increase in *percent_cited*.

Sub-categories

We performed a similar analysis using *Scopus*' sub-categories as predictors of *percent_cited*. As in the analysis above, controlling for open access and impact factor, none of the sub-categories were significantly associated with appearance in Wikipedia; open access, too, was not a significant predictor. Again, the predictor most associated with how frequently a journal is cited in Wikipedia is the journal's impact factor. The effect size of impact factor was qualitatively identical to the analysis above.

Conclusion

Previous research has raised concerns that Wikipedia's use of scientific references is biased toward particular topics and sources that are easily available, such as open-access journals. Such bias, if present, may have major impact as millions of people rely on Wikipedia for high-quality information, including information about science. Most pre-

vious studies examined Wikipedia citations only, without comparing citation practices on Wikipedia to those within a suitable sample of the scientific literature. This study, in contrast, evaluated evidence for bias in representation of science by examining a large swatch of the scientific literature scientists rely on most.

We find that the chief predictor of whether a journal is cited on Wikipedia is its impact factor and general research area. Crucially, whether the journal is or is not open access is not associated with its representation on Wikipedia.

Nevertheless, the present research is beset with a number of limitations that leave important questions unanswered. Perhaps most importantly, the procedure employed to link Wikipedia citations to journals indexed by Scopus successfully identified only about 55% of the citations. It is possible that the bulk of these citations are to sources outside the conventional scientific literature or to very low-impact journals omitted from our data sample. Our qualitative analysis of these items indicates that many of them point to blogs and popular media outlets, e.g. New York Times. The present study cannot address the concern expressed by others, e.g. [9, 10], that sources outside the scientific literature are used too heavily in scientific articles.

Furthermore, while open access does not appear to play a role in the representation of important science in the English-language Wikipedia, it may loom large in promoting access to scientific information (and thus referencing on Wikipedia) in relatively poor countries. Research currently underway will explore the role of open access in referencing on Wikipedia of all major languages.

These findings should be interpreted with care because it is unclear whether Wikipedia or the academic literature should be taken as the golden standard of impact. Impact factor within the academic literature is a notoriously contentious metric, especially across research areas (Seglen, 1997). Thus, inconsistencies between academic and Wikipedia citations may signal that some academic journals are over- or under-cited, rather than over- or under-represented on Wikipedia.

References

Baum, C. (2008). Stata tip 63: Modeling proportions. Stata Journal *8*, 299–303.

Björk, B.-C., and Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. BMC Medicine *10*, 73.

Denning, P., Horning, J., Parnas, D., and Weinstein, L. (2005). Wikipedia Risks. Commun. ACM *48*, 152–152.

Ford, H., Sen, S., Musicant, D.R., and Miller, N. (2013). Getting to the Source: Where Does Wikipedia Get Its Information from? In Proceedings of the 9th International Symposium on Open Collaboration, (New York, NY, USA: ACM), pp. 9:1–9:10.

Lucy Holman Rector (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. Reference Services Review *36*, 7–22.

Luyt, B., and Tan, D. (2010). Improving Wikipedia's credibility: References and citations in a sample of history articles. J. Am. Soc. Inf. Sci. *61*, 715–722.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F.Å., and Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. J Assn Inf Sci Tec *66*(2), 219-245.

Nielsen, F.A. (2007). Scientific citations in Wikipedia. First Monday *12*.

Samoilenko, A., and Yasseri, T. (2014). The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. EPJ Data Science *3*.

Seglen, P.O. (1997). Why the impact factor of journals should not be used for evaluating research. BMJ *314*, 497.

Shuai, X., Jiang, Z., Liu, X., and Bollen, J. (2013). A Comparative Study of Academic and Wikipedia Ranking. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, (New York, NY, USA: ACM), pp. 25–28.

Spoerri, A. (2007). What is popular on Wikipedia and why? First Monday *12*.

"Traffic to Wikipedia." Retrieved January 24, 2015. http://www.alexa.com/siteinfo/wikipedia.org.

"Trends in Articles and Editors." Retrieved Jan. 24, 2015. http://commons.wikimedia.org/wiki/File:WMFArticlesVsContrib.png.

"Verifiability." Retrieved Jan. 24, 2015. http://en.wikipedia.org/wiki/Wikipedia:Verifiability.

"Scopus Overview." Retrieved Jan. 24, 2015. http://www.elsevier.com/online-tools/scopus/content-overview.

"PloSOne Publishes 100000th Article." Retrieved Jan. 24, 2015. http://blogs.plos.org/everyone/2014/06/23/plos-one-publishes-100000th-article/.